Update on Confidentiality and Disclosure Avoidance

Jennifer Hunter Childs and John M. Abowd

Research and Methodology Directorate

U.S. Census Bureau

National Advisory Committee

November 8, 2019

Shape
your future
START HERE >

New Research on Confidentiality

Background

Are respondents worried about re-identification?

Do respondents prefer more privacy at the cost of less accuracy of publicly released data or are they willing to risk privacy for more accurate and useful data?

The terms and concepts are familiar to economists, data scientists, and survey researchers but are not something respondents have thought about.

Methods

In-person cognitive testing, think-aloud with intermittent probing

- Round 1: 27 interviews
- Round 2: 17 interviews

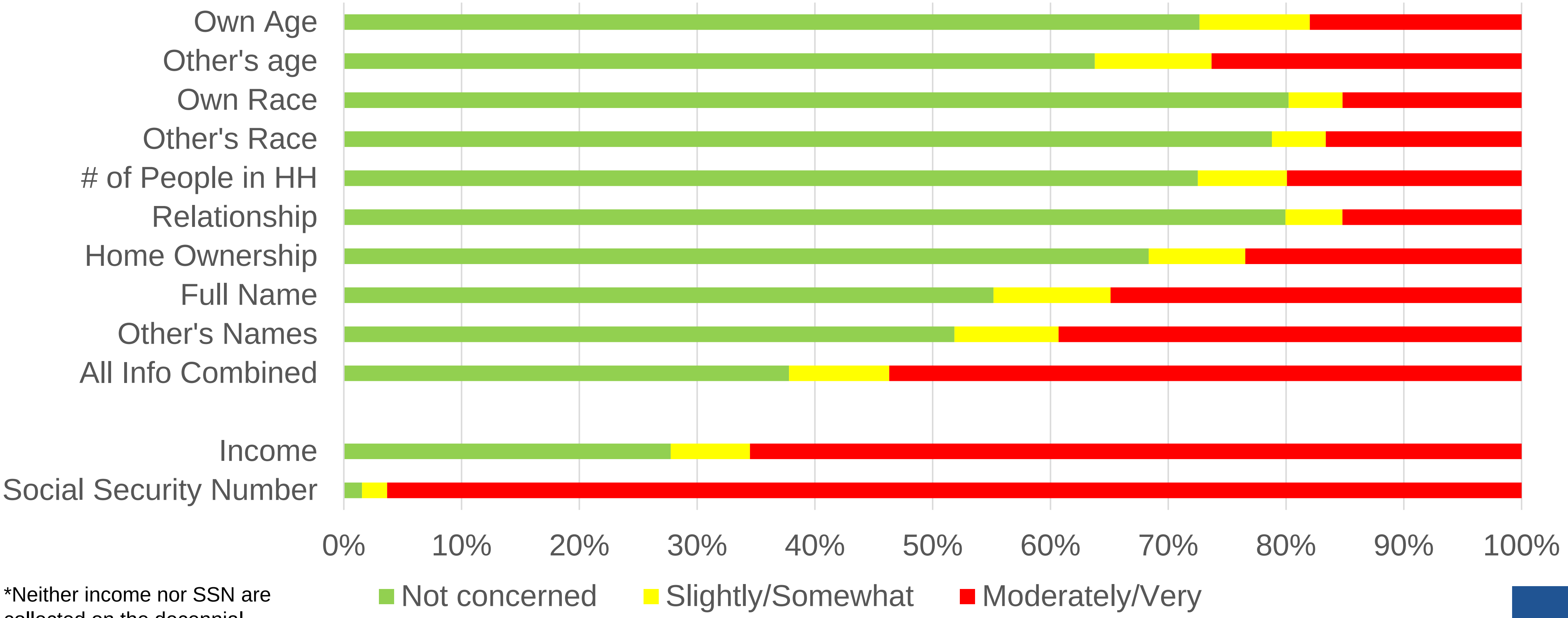
Proof of concept field test

- Online instrument
- Randomized National Sample: 20,000 households, up to 3 emails per household
- 727 responses after cleaning
- Half of sample had web probes

Goal: Large, nationally representative sample survey

Data approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY19-CED002-B0003; CBDRB-FY19-CED001-B0015.).

Level of Concern about Confidentiality of Census Information



*Neither income nor SSN are collected on the decennial census.

Data approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY19-CED002-B0003; CBDRB-FY19-CED001-B0015.).

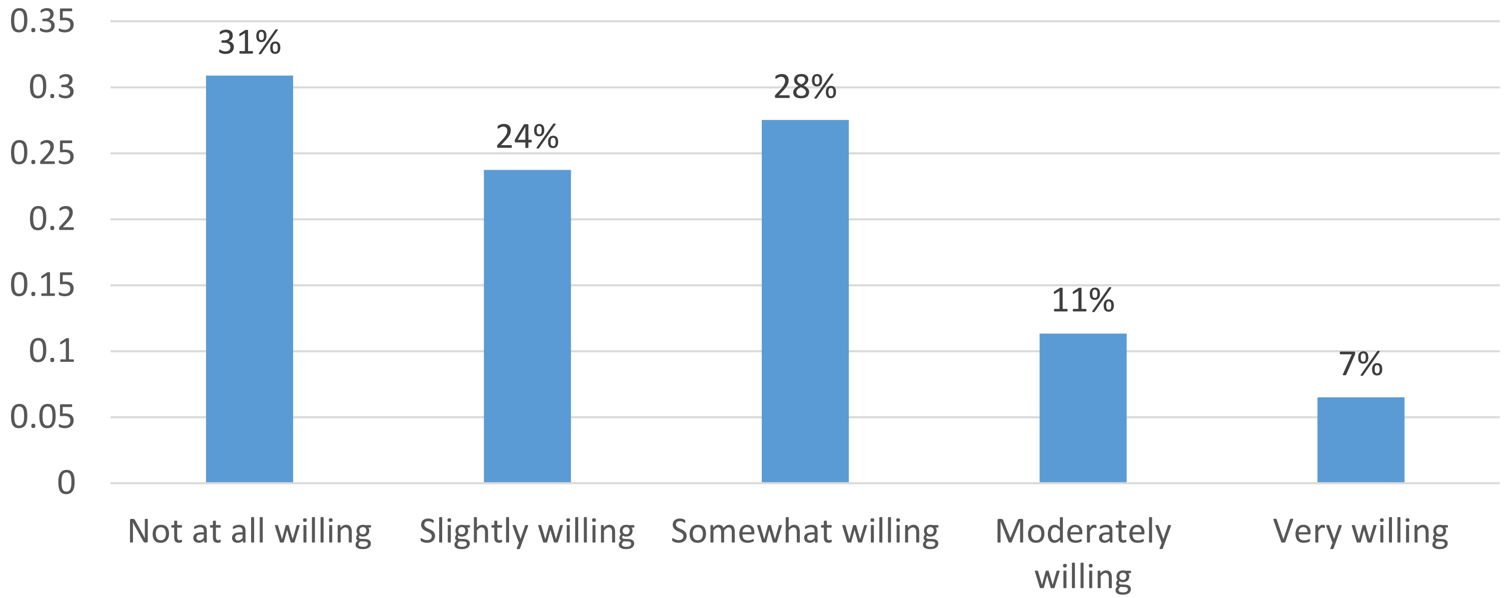
Shape your future START HERE >

United States® Census 2020

Privacy-Accuracy Trade-Off Version 1

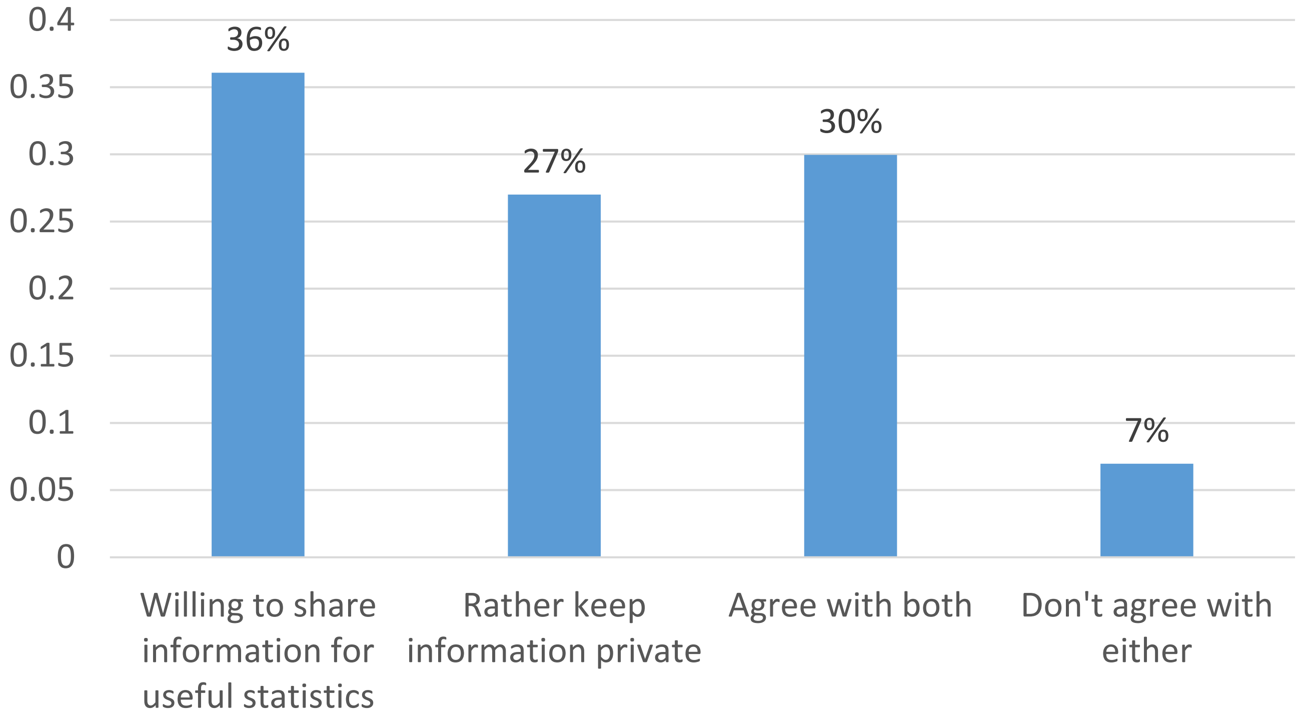
Policy makers, businesses, and researchers use information collected from government surveys to make important decisions. The more detailed the data provided by households like yours, the more useful that information is. This might mean reporting data by ZIP code, for example, instead of by state. But providing more detail may increase the risk that an individual household's information will be identified, even if that risk is low.

In general, how willing are you to risk your confidentiality so the government can produce useful data and statistics for policy makers, businesses and researchers to use?



Data approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY19-CED002-B0003; CBDRB-FY19-CED001-B0015.).

Willingness to Share Information



Data approved for release by the Census Bureau’s Disclosure Avoidance Review Board (CBDRB-FY19-CED002-B0003; CBDRB-FY19-CED001-B0015.

Discussion

Using the term “re-identification” is a problem

- Respondents seemed to understand the behavior and definition but the term was confusing
- Continuing to refine this with additional pretesting

Next steps

- Working on finalizing instrument and sample for larger, more representative study
- Results will inform the 2020 Disclosure Avoidance System

Disclosure Avoidance Update

2018 End-to-End Census Test

Version of the 2020 Disclosure Avoidance System used for 2018 E2E Test code base and draft technical documents released https://github.com/uscensusbureau/census2020-das-e2e

Prototype PL94-171 released from the 2018 E2E Test used a privacy-loss budget of 0.25 for reasons documented here: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/memo-series/2020-memo-2019_13.html

Formal Privacy for the American Community Survey

Formal privacy methods for the American Community Survey *will not be implemented before 2025* (Deputy Director's blog: https://www.census.gov/newsroom/blogs/random-samplings/2019/07/boost-safeguards.html)

The scientific and user communities will be fully engaged as part of that process

All current efforts are focused on formal privacy methods for the 2020 Census

2020 Disclosure Avoidance System (Persons)

Expanded from 2,012 cell national histogram (used for E2E Test) to 370,994 cell histogram for Demographic and Housing Characteristics-Persons (DHC-P)

Person-level workload optimization completed for fully interacted:

- HHGQ (In household or in GQ (7 major types))
- Race (63 OMB categories)
- Hispanic/Latino
- Sex
- Age (single years to age 115)

2020 Disclosure Avoidance System (Housing and Households)

387,072 cell histogram for Demographic and Housing Characteristics-Housing (DHC-H)

Household-level workload optimization completed for fully interacted:

- Householder attributes: [race, Hispanic/Latino, sex, age]
- Multigenerational
- Household size (Top-coded at 7)
- P60, P65, P75 (Presence of people 60 years and over (respectively, 65 or 75))
- Household type attributes

2010 Demonstration Data Products

Based on the national 2010 Census Edited File

Accommodates all six PL94-171 tables approximately 70% of tables in DHC-P and DHC-H

Based on each of the workload-optimized approx. 400,000 cell histograms at each geography using new version of the Census TDA (TopDown Algorithm)

Privacy-loss budget 6 (4 for Persons, 2 for Housing Units)

Released October 29, 2019

CNSTAT workshop to discuss demonstration products December 11-12, 2019

New code base release scheduled for November, 2019

Thank you.

John.Maron.Abowd@census.gov and

Jennifer.Hunter.Childs@census.gov

Exact Questions from CBSM Survey

□ A. I am willing to share information about me and my household with some government agencies (like the Census Bureau) so the government can produce more useful data and statistics, even if it means having less control over that information.

□ B. I would rather keep information about me and my household private even if it means the data and statistics produced by the government are less useful.

□ C. I agree equally with both

□ D. I don't agree with either

Exact Questions from CBSM Survey

As you may know, different government departments and services collect data about individuals, for example your tax records and health records. People have different views about whether this data should be used for new purposes after it has been collected. Using this data can bring benefits, such as finding more effective medical treatments or using information about local communities to plan local schools or roads. But some people worry that other uses for data risk their privacy and security, by linking different types of data together and potentially allowing them to be identified.

Overall, which of the following statements comes closest to your opinion?

□ A. Government should find new ways to use data already collected because it benefits public services and society.

□ B. Government should not use data already collected in new ways due to the risks to people's privacy and security.

□ C. I agree equally with both.

□ D. I don't agree with either.